PRATIK SONTAKKE

AI Engineer | +91 8888648824 | pratikass488@gmail.com | Linkedin | Github | Portfolio | YouTube

Professional Summary

Innovative AI Engineer leveraging a robust backend and cloud engineering background to build and deploy scalable, production-ready AI systems. Expertise in developing RAG applications and AI agents with Python and LangChain, grounded in extensive experience with Java, Spring Boot, and AWS infrastructure.

Technical Skills

AI & Machine Learning: LLMs, RAG, AI Agents, Fine-Tuning (PEFT), LangChain, LangGraph
Backend & API Development: FastAPI, Spring Boot, Microservices, REST APIs
Languages: Python, Java, JavaScript, SQL
Databases: Vector DBs (Chroma, Pinecone), PostgreSQL, MySQL, Redis
Cloud & Infrastructure: AWS, Docker, Kubernetes (k8s), Terraform
CI/CD Deployment: GitHub Actions, Vercel

Projects

Multi-Tenant RAG-as-a-Service Platform | | |

- Engineered a RAG SaaS platform **projected to save clients over $20,000 and 1-2 months in development costs** by providing a ready-to-use, production-grade conversational AI.
- Architected a secure, multi-tenant backend on AWS **to guarantee complete data isolation and build client trust**, using a schema-per-tenant model in PostgreSQL and a fine-grained RBAC system.
- Collaborated on a high-performance FastAPI API, **enabling real-time, context-aware responses that ensure a seamless and interactive user experience.**

Intelligent RAG Assistant with Google Workspace Integration |

- Developed a multi-modal RAG assistant that **boosts user productivity by an estimated 20%** by centralizing knowledge workflows and automating email communication.
- Engineered features that **cut document research time by up to 75%** by ingesting Google Drive files, and **reduced email drafting time from minutes to seconds** by leveraging Google Gmail APIs.
- Orchestrated the end-to-end AI pipeline using n8n to **create a seamless, unified user experience** between OpenAI models, a PostgreSQL/pgvector store, and Google Workspace.

Professional Experience

Cloud & Backend Engineer - Freelance Nov 2024 - Jun 2025 | Remote

- Engineered the core AWS infrastructure using Terraform and established a CI/CD pipeline that **slashed deployment times from hours to under 10 minutes**, guaranteeing consistent and error-free releases.
- Developed and containerized a key microservice in Java and Spring Boot.

Software Engineer - Guenstiger Apr 2023 - Oct 2024 | Delhi, India

- Developed and optimized high-performance REST APIs in Java and Spring Boot, reducing API response times by an average of 200ms.
- Led backend enhancements that improved system throughput by 15%, supporting a growing user base without additional infrastructure costs.

Junior Technical Consultant - Edifition Aug 2021 - Apr 2022 | Remote

- Translated client business needs into detailed technical specifications for 5+ projects, which **led to a 15% reduction in requirement-related change requests** post-development.

Education

AI Engineering
MisogiAI By Masai, Bengaluru, India Jun 2025 - Present

- **Coursework:** Gen-AI, Machine Learning, FastAPI, RAG, AI Agents, MCP Server, Prompt Engineering.

Bachelor of Science (B.Sc), Computer Science
Sant Gadge Baba Amravati University, Maharashtra, India Aug 2018 - Nov 2020